DOCUMENT RESUME

ED 179 579 TM 009 861

AUTHOR Williams, E. Belvin
TITLE Ability Testing and Affirmative Action: Some Theoretical and Practical Considerations.
PUB DATE May 79
NOTE 35p.; Paper presented at the annual meeting of the American Psychiatric Association (Chicago, IL, May 14-18, 1979)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability Identification; Achievement Tests; Aptitude Tests; *Competitive Selection; *Culture Free Tests; Decision Making; Educational Testing; Measurement Goals; *Occupational Tests; Personnel Selection; Racial Differences; Socioeconomic Influences; *Test Bias; *Testing Problems; Test Interpretation; Test Validity

ABSTRACT

Issues related to ability testing and affirmative action are discussed, including the fair use of tests in academic and occupational selection, the effects of testing upon the tested individual, and implications of the issues for the major participants. Types and functions of ability tests, both aptitude and achievement, are described. Studies of test fairness or bias or adverse impact are cited, generally indicating similar patterns of prediction for various ethnic and socioeconomic groups, in spite of differences in mean scores. A number of different models of fair selection are contrasted. The use of test scores by institutions to make decisions about individuals is discussed and contrasted with the possibility of the use of test scores by individuals for the enhancement of self understanding and self respect. The implications of the use of these selection instruments and selection procedures are discussed as they refer to the test maker, the test user, and the test taker. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED179579

U.S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Ability Testing and Affirmative Action:

Some Theoretical and Practical Considerations

E. Belvin Williams

Educational Testing Service

Princeton, New Jersey

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

E Belvin Williams

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

TM009 861

Paper presented at the Annual Meeting of the American Psychiatric Association, 14-18 May 1979, Chicago, Illinois.

Ability Testing and Affirmative Action: Some Theoretical and Practical Considerations

Introduction

I appreciate very much the opportunity to participate in the meeting of the American Psychiatric Association, and am indebted to Dr. Harrison-Ross for organizing this session, "Affirmative Action in Psychiatry."

In considering the issue of ability testing and affirmative action, one must first identify the different types of ability testing--i.e., aptitude and achievement; the variety of contexts wherein such tests are used--e.g., educational, occupational; and the usage models which pertain--e.g., placement, selection, certification.

As there has been much controversy about the selection of applicants based upon the results of ability tests in both academic and occupational settings, some time will be devoted to a brief overview of the "fair selection" literature and some alternatives which have been proposed. Given the general thrust of this session, consideration shall also be given to the impact of ability testing in these settings upon the individual. I will conclude with some observations about the implications of the issues identified for the major participants in this process--the test makers, the test users and the test takers.

The author wishes to express his appreciation to Ms. Audre Griffiths for her valuable assistance in the development of this paper, and to Ms. Monica Thatcher for the preparation of the manuscript.

Types and Functions of Ability Tests

The news media has devoted much attention in recent years to the proliferation of standardized testing. While some critics call for the cessation of such testing, the requirements of a society faced with limited resources, an increasingly underskilled workforce, and a consumer movement calling for greater accountability on the part of business and government point instead to a greater reliance on testing.

However, while ability testing has found itself in the arena of public debate, there is often a lack of common understanding of the basic terms and definitions necessary for constructive consideration of this issue. Thus, permit me to begin with a brief overview of some of the types and functions of ability tests.

Aptitude tests measure reasoning ability not directly related to any specific curriculum and assume a common knowledge base. Achievement tests measure developed ability in a particular subject or discipline (e.g., history, literature), where it is recognized that levels of knowledge or competence may differ. Thus, if a student were given a reading passage about the Battle of Waterloo, and asked to answer questions related to the content of the passage, this would constitute an aptitude test. Were those same questions administered without the reading passage, this would constitute an achievement test.

In academic or educational settings, aptitude tests may indicate the readiness of a student for further or specialized training, or offer

information about the probability of a student's success on the performance criterion in question (e.g., a "gifted" program in an elementary school). This is generally referred to as the selection function. Achievement tests may also be used for admission or selection when identification of an individual's present competence in the domain assessed is necessary to the admission decision (e.g., entry into an intermediate French class). They may also serve a diagnostic purpose. An individual's score on an English composition test might indicate that he or she should receive additional instruction in writing; in addition, the nature of the student's performance on particular sections of the examination might provide valuable information as to his or her proper placement within the instructional sequence. Ability tests may also be used in program evaluation and guidance.

In employment settings, ability tests are used in the selection or licensure process for many occupations, among them police, fire fighters, skilled laborers, insurance brokers, administrators, and lawyers. Achievement tests would be used to measure an individual's knowledge of a particular job or job-related domain (e.g., metallurgy, real estate law), while aptitude tests would be used to determine the probability of success on a criterion such as a training program or a particular job or job function.

The need to assess how much students have learned has been integral to the educational process since there have been students and teachers; employment testing, by comparison, is a relative newcomer. However, with the size of the labor force, the fierce competition for jobs, and the

increasing unreliability of educational credentials in terms of skill level and qualifications represented, managers seeking to make objective and fair judgments in selecting, placing, and promoting employees have turned to testing.

In public hearings held by the National Research Council's Committee on Ability Testing, several professionals in the field of employment testing elaborated upon these issues.

Theodore J. Carron, an industrial psychologist with the American Petroleum Institute, and Chairman of the Committee on Personnel Selection in the Department of Refining observed:

> Many jobs [in the petroleum industry] are associated with hazards of toxicity, fire and explosion... The industry's outstanding safety record is, in a large measure, the result of good selection and training procedures.
>
> ... most of our jobs require extensive formal training and only part of this can be done on the job. Secondly, the issue of safety for inexperienced applicants trying to prove themselves is a very considerable problem. And finally, the sheer numbers of applicants make this an economic impossibility. One colleague in a large company has cited a recent instance of 4,439 applicants for 126 openings for pipefitter jobs at a single refinery. There is no economically feasible way to involve these applicants in job trials, and so there must be some effective mechanism to select one applicant over another. The only fair way to do this is on the basis of ability.

Also of interest is a statement by C. Paul Sparks, Coordinator of Personnel Research at Exxon, on the use of aptitude tests in the workplace:

> The current emphasis on employment of individuals in non-traditional jobs means that many applicants appear with no prior record of training or experience in the area for which employment is being sought. Thus,

> interviews are unable to probe for knowledge or skill in the area involved. In these circumstances an aptitude test result may well be the best evidence that the applicant shows promise as a prospective employee.

The technology that has been created to support testing activities is impressive. Among other things, it enables us to gain reasonably precise estimates of the reliability and validity of information gained. These two attributes are fundamental considerations for any information regardless of how it is collected, be it through interview, on-site observation or physical measurement. In addition, it is possible to design the information collection effort in such a way as to be able to support reasonably well generalizations to domains of knowledge, performance, and skills beyond the scope of the immediately observed. Finally, public repeatability and/or empirical confirmation of the entire process is easily done. These are the essential characteristics of objectivity.

Test Bias

In both educational and occupational settings, there can be no argument that a nuclear set of psychologists, psychometricians, and measurement theorists have been successful in creating instruments with high social utility. However, in some instances they have created problems--institutional problems, social problems, and problems for the individual--which now, more than ever before, are receiving intense public attention. What are some of the problems?

Concern with possible bias in the construction of standardized tests due to cultural factors has been longstanding. Anastasi (1949) discussed some of the problems test constructors face in an assessment milieu where

sex differences, varying socioeconomic levels, and ethnic minorities exist. She asked: "First, to what extent is test performance determined by cultural factors? Second, what shall we do about it?" (p.13). We shall look at each of these questions in turn.

In the last decade, a number of researchers have conducted various studies into the nature and extent of cultural bias in standardized tests. The research evidence clearly shows that in matters of predicting college performance of Black and white students (e.g., first year grade point averages), instruments such as the Scholastic Aptitude Test of the College Board are as accurate in predicting the academic performance of Blacks as they are for whites (Cleary, 1968; Davis & Kerner-Hoeg, 1971), in spite of what some critics of standardized testing have alleged (Clark & Plotkin, 1964). Results similar to these have been found when the Law School Admission Test was used (Linn, 1974).

Angoff and Ford (1973) investigated test bias on the Preliminary Scholastic Aptitude Test (PSAT) using an item by group interaction strategy. Ten samples of Black and white students who took the PSAT in October 1970 were selected.

Comparisons of the correlations of the test item difficulties among the various groups showed considerable disparity between Black and white groups who were selected at random. When the groups were equated on ability, the disparity of item difficulties fell considerably, leaving much less to be accounted for by strictly cultural influences. Further, some small disparities were found even within different Black groups if

they were grouped by city of origin, or by urban versus rural background. All of this would seem to indicate that some of what has been ascribed to differences in culture between the races might more accurately be explained by other, non-ethnic influences.

However, the impact of specific cultures on test performance is seen in the study of Chinese, Jewish, Black, and Puerto Rican first grade students completed by Lesser, Fifer, and Clark (1965), and a replication of the study by Stodolsky and Lesser (1967). A third study reported by Flaugher (1971) using Project Access data was yet another replication of the earlier Lesser et al. work, but with an older age group--namely, students in the 11th grade and with a Mexican-American group rather than a Puerto Rican group. In all three studies high SES students tended to out-perform low SES students. A striking and consistent finding, however, was that each ethnic group repeated its own distinctive pattern of mental abilities across the three studies, indicating that there are in fact some areas of ability in which ethnic groups tend to do better, relative to others. But the overall levels of ability remained tied to socioeconomic status, not ethnicity.

In an occupational setting, a six year study conducted by Campbell et al. (1973), indicated that tests valid for one ethnic group (i.e., Black, Caucasian or Mexican-American) were also valid for either of the remaining two groups. Three occupational groups were used--cartographic technicians, inventory management specialists, and medical technicians. Each of the three ethnic groups was represented in each of

the three occupational categories. Three criteria (i.e., overall ratings, job knowledge tests, and work samples) were used, against which twelve measures of aptitude were compared for predictive effectiveness. In general it was found that tests valid for one type of criterion were also valid for the others.

The above cited studies are not presented as exhaustive, but as exemplary of the breadth of the research interest. Excellent reviews by Cronbach (1975), Flaugher (1974), Linn (1974), Bray and Moses (1972), and Humphreys (1952) are readily available.

While it is true that in the above referenced studies no predictive bias was found, more times than not (with the exception of the studies by Lesser et al.) minorities relative to non-minorities scored lower on the tests. In some legal circles (and nonlegal ones as well) these simple mean differences are regarded as prima facie evidence of test bias. On occasion this kind of difference has been labeled as "adverse impact." But, as Anastasi (1949), Turnbull (1949) and Thorndike (1971) have indicated, this notion of test bias has little, if any, instructive value and can easily be viewed as a proper function of measurement, (i.e., to show differences where in fact they exist). But the problem of test bias and predictive validity does not end here, that is, with or without mean differences between subpopulations. In fact, it is the beginning of a more complex consideration which addresses Anastasi's second question--namely, what to do about it.

Fair Selection

In discussions related to test bias, models of fairness in selection are often considered. Before briefly characterizing some of these, I would like to note that the more accurate term for this aspect of the issue is selection bias. The distinction is a useful one in that it emphasizes the fact that these models would apply even if a predictor variable other than a paper and pencil test were employed (e.g., structured interview, grade point average).

A variety of fair selection models, each, as noted by Flaugher (1978), reflecting a particular value system, has been developed by psychometricians to identify and adjust for bias in the selection process. Cole (1973) identified six such models:

> The Quota Model: The quota model of bias involves the idea that fairness lies in some specified proportional representation (p. 237).
>
> The Regression Model: The regression model of test bias follows from definitions of bias which deal with consistent errors of prediction. For example Cleary considered the use of a test to be unfair if it consistently over or under predicted criterion scores for members of some group (p. 238).
>
> The Subjective Regression Model: Darlington (1971) argued that fairness can only be achieved by a kind of combination of the regression model and the type of value judgments made in the quota mode (p. 238).
>
> The Equal Risk Model: Another definition of bias had led to a different selection model. Guion (1966) stated that: "Unfair discrimination exists when persons with equal probabilities of success on the job have unequal probabilities of being hired for the job [p. 26] (pp. 238-239).
>
> The Constant Ratio Model: Thorndike (1971) proposed yet a fifth definition of bias or its complement,

> fairness. In a fair selection procedure, the qualifying scores on a test should be set at levels that will qualify applicants in the two groups in proportion to the fraction of the two groups reaching a specified level of criterion performance [p. 63] (p. 239).
>
> The Conditional Probability Model: The basic principle of the conditional probability selection model is that for both minority and majority groups whose members can achieve a satisfactory criterion score (Y > Yp) there should be the same probability of selection regardless of group membership (p. 240).

Petersen and Novick took the position that these models were inadequate to their task and that more complex analyses were required. In 1976 they introduced a new model for fairness--the Threshold Utility Model. In "An Evaluation of Some Models for Culture-Fair Selection" they state:

> There exists a body of quantitative reasoning, whose origins are ancient and remote, that has received codification in this century in the work of Von Neumann and Morgenstern (1947), Wald (1950), and others. In the theory of the rational-economic man, developed in these writings, when all probabilities of outcomes are assumed known ... there is a simple paradigm required for rational decision. In that paradigm the desirability or utility of each possible outcome is stated quantitatively. Then, given all available information concerning the person in question, the probability of each possible outcome is stated for each decision under consideration. Next, for each possible decision, the utility of each outcome is multiplied by the probability of each outcome and the products are summed to provide an expected utility. Finally, that decision is then made for which the expected utility is highest (p. 25).

In the Threshold Utility Model, the underlying value system is that preference may be given to a particular group or groups, depending upon the utility indicated by existing socially (or other) desired ends. As Johnson (1978) explains:

> If decision-makers felt that it was useful and important to have fair numbers of Black and female physicians, then a utility function would be chosen for use in selection. This function would give weight to the minority or sex status of applicants. A different utility function would be constructed if other priorities were present.

As indicated by Petersen and Novick, and others commenting upon the Threshold Utility Model (Johnson, 1978; Flaugher 1978), this model provides for an explicit statement of utility for each subpopulation, thus permitting public debate on the appropriateness of the established utilities among interested and relevant parties.

What this discussion brings full circle is the continuing quest for a working relationship between theory and practice, and research and "the real world." As noted by Flaugher (1978),

> Novick and Petersen (1976) have made significant strides toward developing a practical system that...is within reach of the statistical prowess of selection practioners. No functioning system in a selection setting is known to this author, however, and probably waits for the packaging of the computerization of the complex statistics involved.

Meanwhile, decisions are being made daily as to which employee shall receive apprenticeship training or which applicant shall be admitted to law school.

What are practitioners to do while the appropriateness of the decision-theoretic model advanced by Petersen and Novick is considered within professional circles and practical usage awaits, among other things, the outcome of this debate and the development of needed computer technology and definitions of disadvantagement?

The internal conflict and debate which has characterized the fair selection issue in psychometric circles is no less evident among those practitioners who must make selections and those members of society who participate in or observe this process. I will briefly discuss two selection strategies which have been advanced and, in some respects, are or have been utilized in various settings.

In a report to the Carnegie Council on Policy Studies in Higher Education, Manning (1977) discussed a two-phase process wherein the first step involved identifying those individuals meeting a minimal standard of admissibility, while the second addressed selection from among these qualified applicants. In the second stage, (not unlike the Threshold Utility Model) Manning would have considered attributes related to the objectives of the institution and the needs of the profession, such as persistence, unusual personal experience, nonacademic interests, career goals, and evaluations by teachers, peers, and relevant others. This is, in effect, the selection process utilized by several institutions of higher education, among them, Harvard.

Lerner (1977) would also establish a pool of qualified applicants who meet some minimal standard (she suggests using current assessment instruments "to separate the qualified from the unqualified") but would then employ random selection (sometimes referred to as the lottery system) to identify the successful applicants. This she sees as preferable to "using the same unvalidated, second-stage screening methods on all qualified applicants...". This "qualified or not" model is common to

trade unions, where all members may be expected to meet a minimum set of qualifications, and relative degrees of qualification beyond that minimum are not germane.

While these models do not resolve the issue of fair selection at the first stage of the process they do have the similarity of placing in clearer relief the distinction between a qualified applicant who is not selected and one who is not admissible. The availability of qualified applicants is often seriously under-estimated--especially by a rejected individual.

A cardinal principle to keep in mind is that we must enable a widening set of alternatives or decision rules to become a legitimate part of the selection process. Where these may be in conflict, limits must be established to permit their coexistence. For example, in the petroleum industry, where gradations of performance may be of great significance due to the high cost--both financial and safety--of poor selection, a limited number of minimally or "less" qualified individuals might be exposed to some training sequence prior to a second screening stage, so that additional data regarding their predicted performance on the criterion in question might be obtained and considered.

The Individual and the Process

It is ironic that a strong historical interest of early workers in educational measurement focused on individual differences. For some educators, the search for and identification of dimensions along which individuals could be ordered was justified by the promise that more

efficient planning or placement could be achieved for each person. But rarely does one encounter in the psychological literature claims to the effect that tests results could be and should be used by the test taker for the enhancement of self-understanding and self-respect. More typically, test data were, and still are, turned over to institutional decision makers for the purpose of achieving institutional objectives.

On this latter point, one is strongly inclined to say that, in our haste to respond to the institutional requests of education, government and industry in a socially efficient way, the well-being of the individual has been subordinated.

A definition of self-respect not alien to those in the health fields has been formulated by Rawls (1971):

> ...self-respect implies a confidence in one's ability, so far as it is within one's power, to fulfill one's intentions. When we feel that our plans are of little value, we cannot pursue them with pleasure or take delight in their execution. Nor plagued by failure and self-doubt can we continue in our endeavors (p. 440).

If we in the educational measurement community were to ask ourselves what is it that we do to enhance a sense of self-respect for each individual who takes our tests, a ready answer would not be easy. It is clear that we assist institutions in maintaining their institutional respect, since many of them point with pride to the test scores of their applicants and their admitted cohorts. But, about what can the individual feel proud or good, even for a moment, for having taken one of our tests? More often than not, self-doubts are raised by incremental failures to answer

successive items satisfactorily. We need to give more to the individual, to treat him or her as an end and not solely as a means to institutiona "gratification."

In addition to considering the problems of those who have made it to the selection stage of the system, it is equally important to concern ourselves with those who are unable to progress this far because of educational failure, or who are unable to find or hold a job. At this stage of the process, the problem lies in assuring individuals a foundation in basic competencies so that there are increased educational and occupational options available.

Although many members of minority communities have grave misgivings about the introduction of standards in the schools and/or basic skills assessment programs, Black leaders are increasingly supportive of such efforts. For example, in a hearing before the Education, Arts, and Humanities Subcommittee of a U.S. Congressional Committee on Human Resources, the Reverend Jesse Jackson noted that he favored standardized testing for basic skills beginning in the first grade "not only to diagnose, but also to get kids used to taking tests."

Similarly, Gustav Heningburg, President and Chief Executive Officer of the Greater Newark Urban Coalition, Inc., observed that "people in minority communities are conditioned to respond negatively to tests of any kind." He is convinced, however, that basic skills testing and the installation of standards in the schools are necessary to enable

minorities to break the cycle of systematic educational failure and chronic unemployment.

Implications for the Participants in the Testing Process

The selection procedures employed by test users are coming under heightened scrutiny as students, applicants, and others seek to assure themselves due process in the testing situation.

The recent introduction of truth-in-testing bills in both federal and state legislatures (the latter including California, New York, and Florida, among others) attests to the increased public concern about the nature of the tests, the testing industry, test use, and the effect of this process on the individual. To date, California is the only state in which such a bill has been enacted. Also, consumer advocate Ralph Nader has expressed much interest in the testing industry--its operation, its accountability, and its influence on social mobility.

As alluded to earlier, societal pressures such as the state of the economy, the increasing participation of females and minorities in professional fields, and the need to assure the job relevance of occupational assessment instruments (occasioned by the 1964 Civil Rights Act, the Equal Employment Opportunity Commission Guidelines on Employee Selection Procedures, and the Griggs decision) have, in fact, resulted in an increase in the testing of individuals as well as in the different kinds of assessment instruments that are utilized (e.g., basic skills, personal characteristics, scholastic aptitude).

Fundamental to the rise in popularity of both psychological and educational tests has been the heightened "decision demand." As expressed by Thorndike and Hagen (1961)

> the educator or the practical psychologist is continually faced with the necessity of arriving at some decision as to a course of action...
>
> Our basic assumption is that sound decisions arise out of relevant knowledge of the individual or individuals.
>
> We assume basically that knowledge is good, that knowledge is power, that knowledge is the basis for effective control of the problems that confront us from day to day. This is a basic tenet of our faith (pp. 7-8).

Accepting the basic tenet of faith, one must ask how it is ascertained that relevant knowledge has been gained. With respect to tests, it is usually said that the confirmation of relevance is gained through the validation of the test. Commenting on this phrase Cronbach (1971) argues that

> One validates, not a test, but an interpretation of a data arising from a specified procedure. Because every interpretation has its own degree of validity, one can never reach the simple conclusion that a particular test is valid (p. 447).

Given Cronbach's formulation, it is possible to conceive of situations in which judgments based on interpretations of data may be misapplied, should two or more specific validation procedures be confused. Meehl and Rosen demonstrated the disutility of test use in that situation in which the occurrence of the phenomenon, (e.g., poor adjustment, good adjustment or college success in the real world), differed markedly from the test's ability to detect or reflect the phenomenon faithfully. Indeed, fewer

correct classification decisions could be made with the test than without it.

Even if the tests were used with proper caution, the problem of fair test use as identified by Thorndike (1971) and Darlington (1971) would still remain. As pointed out by both of these authors and further articulated by Gross and Su (1975), and Petersen and Novick (1976), integral to the technical problem is a value decision that must be made by someone on someone's behalf and probably imperfectly, regardless of the amount of data or information available.

The problem of test misuse is inextricably intertwined with specific technical limitations of tests, and the general limits of all empirical inquiry. Within professional psychology there has been a general acknowledgement of a responsibility to guard against over reliance upon test data and a misuse of it. The professional interest in test construction, administration and interpretation spans almost a half century. Recognized and widely accepted guidelines regarding the construction and publication of tests were developed almost a quarter of a century ago (American Psychological Association, 1954) and have received periodic review and revision (American Psychological Association, 1966, 1974).

Clearly, test publishers have particular responsibilities in such efforts. They should set forth clearly and precisely the limitations of their instruments. At the very minimum, the burden of providing relevant information about testing objectives, validation strategies, and interpretive constraints lies with test publishers for their particular

instruments. Further, test makers should provide complete and straightforward information to the test taker about the nature of the test and the meaning of his or her scores. Finally, there must be almost unilateral effort on the part of test publishers to investigate possible sources of error that are both controllable and uncontrollable. Professional groups--be they psychiatrists, educators, lawyers or others--certainly must critically question in constructive and knowledgeable ways the uses and possible misuses of test instruments.

To this point it has been argued that we face profound social problems, that psychology has contributed to the development of certain of these problems, and to some extent has taken steps to correct the imbalance and troublesome circumstances. But, the well-intentioned social steps have been insufficient to quiet the larger social concern.

Churchman (1973) aptly stated the necessity of giving foremost consideration to the systematic and pervasive impact of technology--in this instance, testing and information processing technologies--on human beings:

> Because morality is so elusive a concept, the careful researcher may be inclined to exclude it from his considerations, arguing that it's largely a matter of personal or cultural style. But I believe we are living through an age when morality is becoming a stronger and stronger force in the whole world, and for us to ignore it in our decision theory research is for us to become largely irrelevant (p. 4).

The near tragedy at the Three Mile Island nuclear plant underscored the need for morality to be the underpinning of all actions having significant

social impact. In our daily use of test scores and other assessment data to make decisions regarding selection, placement, and certification, we must be constantly mindful of the impact of our efforts on the life choices--and in some cases, life direction--of the individuals involved. In his address of a decade ago, Professor Miller argued:

> The most urgent problems of our world today are the problems we have made ourselves. They have not been caused by some needless or malicious inanimate Nature, nor have they been imposed on us as punishment by the will of God. They are human problems whose solutions will require us to change our behavior and our social institutions (1969, p. 1063).

The assessment and evaluation of human behavior is inevitably embedded in the very fabric of our existence and permeates all that we do. Tests are but a small part of the undertaking. What is needed is renewed commitment to the understanding of people as human beings supported by theoretical constructs and documented fact gathered in a variety of different ways, as well as commitment to the changes that may be called for as a result of that understanding. Whatever the empirical strategies that might be adopted or discovered, they must always enable us to assess the fallibility of our ways. Such attributes are not unique to tests alone, but are fundamental to any mode of open and full inquiry. Made richer by the expanded participation of all those who seek to know, the self corrective mechanism of scientific inquiry as a social enterprise is probably the only safeguard against test bias and selection unfairness. This means that no claim is beyond doubt and review, and all claims deserve a hearing.

REFERENCES

American Psychological Association. Standards for educational and psychological tests. Washington: APA, 1974.

American Psychological Association. Standards for educational and psychological tests and manuals. Washington: APA, 1966.

American Psychological Association, American Educational Research Association, and National Council on Measurement Used in Education. Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 1954, 51. (Supplement)

Anastasi, A. Educational Measurement & the Law: Proceedings of the 1977 ETS Invitational Conference. Princeton, N.J.: Educational Testing Service, 1977.

Angoff, W.H., and Ford, S.F. Item-race interaction on a test of scholastic aptitude. Journal of Educational Measurement, 1973, 10, 95-106.

Bray, D.W., and Moses, J.L. Personnel selection. Annual Review of Psychology, 1972, 23, 545-576.

Campbell, J.T., Crooks, L.A., Mahoney, M.H., and Rock, D.A. An Investigation of sources of bias in the prediction of job performance. A six year study. Final Project Report (PR-73-37). Princeton, NJ: Educational Testing Service, 1973.

Carron, T.J. Statement to the Committee on Ability Testing of the National Research Council. Washington, D.C., November, 1978.

Churchman, C.W. Morality as a value criterion. In J.R. Cochran and M. Zeleny (Eds.), Multiple criteria decision making. Columbia, SC: University of South Carolina Press, 1973.

Clark, D.B., and Plotkin, L. The Negro student at integrated colleges. New York: National Scholarship Service for Negro Students, 1964.

Cleary, T.A. Test bias: Prediction of grades of Negro and white students in integrated colleges. Journal of Educational Measurement, 1968, 5, 114-124.

Cole, N.S. Bias in selection. Journal of Educational Measurement, 1973, 10, 237-255.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.) Educational Measurement. Washington: American Council on Education, 1971.

Cronbach, L.J. Five decades of public controversy over mental testing. American Psychologist, 1975, 30, 1-14.

Darlington, R.B. Another look at "culture fairness." Journal of Educational Measurement, 1971, 8, 71-82.

Davis, J.A. and Kerner-Hoeg, S. Validity of pre-admissions indices for blacks and whites in six traditionally white public universities in North Carolina (PR-71-15). Princeton, NJ: Educational Testing Service, 1971.

Flaugher, R.L. Bias in testing: A review and discussion. TM Report 36. Princeton, NJ: ERIC Clearinghouse on Tests, Measurement, and Evaluation, 1974.

Flaugher, R.L. Current Issues in Fair Testing. A paper prepared for the Massachusetts Division of Personnel Administration, December 1978.

Flaugher, R.L. The many definitions of test bias. American Psychologist, 1978, 33, (7), 671-679.

Gross, A.L. and Su, W. Defining a "fair" or "unbiased" selection model: a question of utilities. Journal of Applied Psychology, 1975, 60, 345-351.

Heningburg, G . President and Chief Executive Officer of the Greater Newark Urban Coalition, Inc. Personal communication, 1979.

Humphreys, L.G . Individual differences. Annual Review of Psychology, 1952, 3, 131-150.

Jackson, J. Chicago Sun Times, April 10, 1979.

Johnson, S. The "Measurement Mystique": Issues in Selection for Professional Schools and Employment. Paper presented to the Committee on Ability Testing Open Meetings, National Research Council, November 1978.

Lerner, B. Equal Protection and External Screening: Davis, De Funis, & Bakke. Educational Measurement & the Law: Proceedings of the 1977 ETS Invitational Conference. Princeton, NJ: Educational Testing Service,1977.

Linn, R L. Test bias and the prediction of grades in law school. Unpublished paper prepared for the Conference on the Future of Law School Admission Council Research, Princeton, NJ: Educational Testing Service, 1974.

Manning, W. H. The Pursuit of Fairness in Admissions to Higher Education. Selective Admissions in Higher Education: A Report of the Carnegie Council on Policy Studies in Higher Education. San Francisco, CA: Jossey-Bass Publishers, 1977.

Meehl, R.E and Rosen, A. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. Psychological Bulletin, 1955, 52, 194-216.

Miller, G.A. Psychology as a means of promoting human welfare. American Psychologist, 1969, 24.

Novick, M.R. and Petersen, N.S. An Evaluation of some models for culture-fair selection. Journal of Educational Measurement, 1976, 13, (1), 3-41.

Rawls, J.A. A theory of justice. Cambridge, Massachusetts: Harvard University Press, 1971.

Sparks, C.P. Statement to the Committee on Ability Testing of the National Research Council, Washington, D.C.: November 1978.

Thorndike, R.L. Concepts of culture-fairness. Journal of Educational Measurement, 1971, 8, (2), 63-70.

Thorndike, R.L. and Hagen, E. Measurement and Evaluation in psychology and education. New York: Wiley, 1961.

Turnbull, W.W. Influence of cultural background on predictive test scores. Proceedings of the 1949 Invitational Conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1949.